# SOME PROBLEMS IN STATISTICAL PATTERN RECOGNITION

by

Somesh Das Gupta*

University of Minnesota

January, 1976

*This work was supported by U.S. Army Research Grant DAAG-29-76-G-0038 at the University of Minnesota.

## 1. Introduction

In this paper we shall review some of the recent work on nonparametric and sequential rules in statistical pattern recognition along with criticisms, and indicate some new results in these areas. Summary reviews of the literature are given by Das Gupta [10] and Kanal [22].

The basic problem in statistical pattern recognition can be formulated as follows. Let $\omega$ be a point in a sample space with an associated $\sigma$-field of events and a probability measure. Let $X(\omega)$ denote a real-valued vector of measurements on $\omega$ and let $I(\omega)$ denote the pattern-class of $\omega$ which takes values in $(1,2,\ldots,k)$. The problem is to predict $I(\omega)$ from the knowledge of $X(\omega)$. Denote the pattern-class probabilities by $\xi = (\xi_1,\ldots,\xi_k)$, where $\xi_i = \Pr(I=i)$, and the class distributions by $F = (F_1,\ldots,F_k)$, where $F_i$ is the conditional distribution of $X$, given $I = i$; let $F_i$ admit a density $f_i$ with respect to a $\sigma$-finite measure $\mu$.

The above problem can arise in different forms due to different situations and structures of available knowledge as discussed below. The problem may be to classify a single unit or more than one unit which may occur in a single batch or in a sequence. Moreover, the units to be classified may belong to the same pattern-class (i.e., when units are sampled from the space conditioned on some value of $I$), or to different pattern-classes. In almost all the problems, $F$ and $\xi$ are not completely known although it may be known that $F_1 \times \ldots \times F_k$ belong to a given set $\Omega$. In order to get more information on $F$ and $\xi$, data in the name of a "training sample" are collected in one of the following ways.

(a) Separate samples from $k$ pattern-class populations, denoted by $\pi_1,\ldots,\pi_k$. In this case $\xi$ is interpreted as a prior probability vector.

(b) Sample from the mixed population (denoted by $\pi$) which is a mixture of $\pi_1,\ldots,\pi_k$ in the proportion $\xi_1,\ldots,\xi_k$.

Sometimes when the units to be classified occur in a sequence, observations on the first $i$ units are taken as a training sample to predict the pattern-class of the $(i+1)$th unit.

A training sample may be identified, i.e., the observations are available on both $X$ and $I$, or unidentified when the observations are available on $X$.

The so-called nonparametric methods arise when $F_i$'s are not given explicitly in simple parametric forms. It is well known that there is no distribution-free rule for this problem. So the performance of a rule cannot be evaluated (except some asymptotic results and broad bounds) without additional knowledge of the underlying $\Omega$. A Bayes-rule can be easily derived when $\xi$ and $f_i$'s are known. The major bulk of the literature is devoted on plug-in versions of a Bayes' rule, i.e., when the unknowns are replaced by their respective estimates (derived from the training sample) in the form of the Bayes' rule. For this nonparametric problem, generally estimates of $f_i$'s and $\xi$ are used; asymptotic properties of such rules are then easily derived from the asymptotic properties of the estimates used. Rules based on tolerance regions and nearest-neighbors are also discussed in the literature; some of these rules indirectly use estimates of density functions. Another class of rules are suggested following the nonparametric methods for the two-sample

problem; these rules are based on general U-statistics, estimates of c.d.f.'s, ranks, and permutation-invariance. A class of rules, called "empirical best-of-class rules" is also under study; these rules are optimum in some sense when they are applied on the identified training sample.

Sequential rules also arise in different situations. There can be rules based on sequential experimentation on the components of $X(\omega)$, although there is not any result in this area worth mentioning (except possibly some heuristic methods). Next, a training sample may be obtained sequentially and rules may be devised based on such sequential experiments. Furthermore, when the units to be classified belong to the same pattern-class, they may also be observed through a sequential experiment. It may be noted that when the units to be classified occur in a sequence from $\Pi$, a sequential rule may be devised, although a sequential experimentation in such a case is not meaningful. All the papers in this area deal with direct applications of sequential two-sample tests. Unfortunately, very little has been done so far.

Although there are some papers on Monte-Carlo studies on performances of some rules, the studies on robustness and relative efficiency have to be done much more intensively and carefully. Asymptotic results are of theoretical interest; however, good bounds on error-probabilities and studies on errors of approximation will be more valuable.

## 2. Notations and Preliminaries.

Let us first restrict our attention to the case $k = 2$, and the situation where the pattern class of one unit is to be predicted. A

decision rule, not depending on the training sample, is given by $\delta = (\delta_1, \delta_2)$, where $\delta_i(x)$ is the probability of deciding $i$ as the correct pattern class, given the observation $X = x$. The probabilities of error for the rule $\delta$ are given by

$$\alpha_1(\delta;f) = \int \delta_2 f_1 d\mu, \quad \alpha_2(\delta;f) = \int \delta_1 f_2 d\mu$$

where $f = (f_1, f_2)$. Then the risk with the prior distribution $\xi = (\xi_1, \xi_2)$ and 0-1 loss function (or, the total probability of error) is given by

$$R(\delta;\xi,f) = \xi_1 \alpha_1(\delta;f) + \xi_2 \alpha_2(\delta;f).$$

Given $\xi$ and $f$, a Bayes rule $\delta^*(.;\xi,f)$ which minimizes $R(\delta;\xi,f)$ is given by

$$\delta_1^*(x;\xi,f) = \begin{cases} 1, & \text{if } \xi_1 f_1(x) > \xi_2 f_2(x) \\ 0, & \text{if } \xi_1 f_1(x) < \xi_2 f_2(x) \end{cases}$$

Let $R^*(\xi,f) \equiv R(\delta^*(.;\xi,f);\xi,f)$.

Let $\theta$ stand for $(\xi,f)$ or $(\xi,F)$. When $\theta$ is known a "good" rule generally depends on $\theta$, and it is denoted by $\delta(.;\theta)$ ; $\delta^*$ is such a rule. We shall drop $\xi$ from $\theta$ when the population is not mixed and $\xi$ is not given. When $\delta$ explicitly depends on $\theta$, we shall write $\alpha_i(\delta;f)$ as $\alpha_i(\delta;\theta)$.

Now, consider the problem when $\theta$ is not completely known. Information on $\theta$ is based on a training sample $S_N$; in case the sampling is done separately from $\pi_i$'s, $N$ will stand for the vector of sample sizes. A decision rule, in that case, is denoted by $\delta_N(.,S_N) = (\delta_{N1}(.,S_N), \delta_{N2}(.,S_N))$. In particular, such a rule may be a plug-in version of $\delta(.;\theta)$, given by $\delta(\cdot;\hat{\theta}_N) \equiv \delta_N(\cdot;\hat{\theta}_N)$, where $\hat{\theta}_N$ is generally chosen to be a consistent estimate of $\theta$; we shall often write $\delta_N(\cdot;\hat{\theta}_N)$ by $\hat{\delta}_N = (\hat{\delta}_{N1}, \hat{\delta}_{N2})$.

The conditional probabilities of error and the conditional risk of $\delta_N(.,S_N)$, given $S_N$, are given by

$$\alpha_{ic}(\delta_N;S_N,f) \equiv \int \delta_{Nj}(x;S_N)f_i(x)d\mu(x), \ (i \neq j), \text{ and}$$

$$R_c(\delta_N;S_N,\theta) \equiv \sum_{i=1}^{2} \xi_i \alpha_{ic}(\delta_N(.,S_N);f).$$

The unconditional probabilities of error and the risk of $\delta_N$ are given by

$$\alpha_i(\delta_N;\theta) = E_N\left[\alpha_{ic}(\delta_N;S_N,f)\right]$$

$$R(\delta_N;\theta) = E_N\left[R_c(\delta_N;S_N,\theta)\right]$$

where $E_N$ denote the expectation over $S_N$.

When there are more than one unit to be classified, and it is known that they come from the same population, the above development can easily be extended. However, when the units occur in a sequence one may adopt Samuel's approach, although no results are available in the literature when the densities are not known. When the units to be classified arise from the mixed population one may use the compound decision approach as suggested by Robbins [34] and later developed by Hannan and Robbins [18], Samuel [36,37] and Yao [53]; however, all these papers assume that the class-densities are known. See VanRyzin [43] for a similar development when the distributions are unknown. When the units to be classified arise in a sequence from $\pi$ and for classifying the ith unit the observations on all the previous identified (by a "teacher") units are used as a training sample, a completely separate theoretical development would be called for. However, all the papers dealing with this problem put emphasis only on the prediction of the class of ith unit using the standard theory discussed above.

## 3. Nonparametric Rules.

### 3.1 A Simple Approach.

Most of the papers deal with asymptotic properties of rules based on a

training sample $S_N$, as $N \to \infty$. In particular, the convergences (and their rates) of conditional, as well as unconditional, probabilities of error and risk are dealt with. Special emphasis is given to the rule $\hat{\delta}_N^*$, the plug-in version of a Bayes rule $\delta^*$. Bounds on probabilities of correct classification for some heuristic rules are also available in some papers. The above convergences as $N \to \infty$ and the number of units to be classified tend to $\infty$ are also discussed in some special cases.

We present the following asymptotic results which can be proved under very general conditions as the size (or sizes) $N$ of the training sample tend to $\infty$.

Suppose $\delta(x;\hat{\theta}_N) \to \delta(x;\theta)$ in probability (a.s.) for almost all ($\mu$) x. Then

(I) $\alpha_{ic}(\hat{\delta}_N; S_N, f) \to \alpha_i(\delta;\theta)$, in prob. (a.s.),

(II) $\alpha_i(\hat{\delta}_N;\theta) \to \alpha_i(\delta;\theta)$,

(III) $R_c(\hat{\delta}_N;S_N,\theta) \to R(\delta;\theta)$, in prob. (a.s.)

(IV) $R(\delta_N;\theta) \to R(\delta;\theta)$.

In the above result for convergences of risks, the condition "for almost all x" may be relaxed by the condition "for almost all x in the set $\{x: \xi_1 f_1(x) \neq \xi_2 f_2(x)\}$."

In particular suppose $\delta_1(x,\theta) = 1$, if $D(x,\theta) > 0$, where $D(x;\theta)$ is a function of $x$ when $\theta$ is known. Moreover, assume that $D(x,\theta)$ equals zero on a null set, (although this condition can be slightly relaxed). If $D(x,\hat{\theta}_N) \to D(x,\theta)$ in probability (a.s.) for almost all x, the above results on convergences hold. The primary requirements for the above conditions to hold are that $D(x,\theta)$ is continuous in $\theta$ for almost all x,

and $\hat{f}_i(x) \to f(x)$ in prob. (a.s.), where $\hat{f}_i$ is an estimate of $f_i$. The detailed proof of these reults will be given elsewhere. For similar but weaker results, see Johns [21] and Glick [16]. It may be noted that the above results do not require $\delta$ to be a Bayes rule and $\hat{f}_i(x)$ integrable; these assumptions are used in almost all the papers in this area. The above results can be used to simplify the proofs of many known results.

The problem may also be handled from decision theoretic-viewpoint with provisions for "withholding decision" or "doubtful regions." See Rao [32] for this approach when the distributions are known and Patrick's book [29] for some heuristic developments when the distributions are unknown.

3.2. Rules based on estimates of density functions.

All the important papers deal with asymptotic properties of $\hat{\delta}_N^*$ with various estimates of $\theta$. Recall that one may define $\delta^*$ in either of the following ways:

A. $\delta_1^*(x;\theta) = 1$ , iff $D(x,\theta) \equiv \xi_1 f_1(x) - \xi_2 f_2(x) \geq 0$ .

B. $\delta_1^*(x;\theta) = 1$ , iff $D_1(x;\theta) \equiv \xi_1 f_1(x)/[\xi_1 f_1(x) + \xi_2 f_2(x)] \geq 1/2$ .

The following methods for estimating $f_i$'s are mostly used.

(i) Fix-Hodges' [12] method; later, modified by Loftsgaarden and Quesenberry [25].

(ii) Aizerman-Braverman-Rozonoer's method [1].

(iii) Čencov's method [6]

(iv) Parzen-Cacoullos' method [28,5].

(v) Wolverton-Wagner-Yamato's recursive method [50,52].

The known results on $\hat{f}_i$ are then used to derive asymptotics for $\hat{\delta}_N^*$ .

Fix and Hodges [12] essentially proved (II) with their suggested estimators of $f_i$'s , when the training sample is separately drawn from $\pi_i$'s and $\xi = (1/2,1/2)$ . Johns [21] proved the same result with a minor modification of Fix-Hodges' estimates when the training sample is identified and drawn from the mixed population. However, Johns considered the problem with a general loss structure and more general space of values for the pattern-class indicator variable I. Van Ryzin [44] proved the result III (in probability) with estimates (ii), (iii) and (iv), when the training sample is separately drawn from $\pi_i$'s . He also studied the rates of these convergences. Van Ryzin [42] proved the result IV with estimates similar to (iv) when the training sample is an identified sample from the mixed population; he also obtained bounds on $R(\hat{\delta}_N^*;\theta) - R(\delta^*;\theta)$ under additional conditions on $\theta$ .

Another approach used in the literature is to estimate $D(x;\theta)$ or $D_1(x,\theta)$ directly. Recursive decision rules are considered (for easier updating), especially when the units to be identified occur in a sequence from $\pi$ and the correct pattern-class of a unit is known after its prediction. For classifying the ith unit, all the observations on the previous units constitute a training sample. Suppose $\hat{\delta}_N^*$ depends on $D_{N1}(x;S_N)$ or $D_N(x;S_N)$ as $\delta^*$ depends on $D_1(x;\theta)$ or $D(x,\theta)$ , respectively; that is, $D_{N1}$ and $D_N$ are respective estimates of $D_1$ and $D$ . It is shown by Van Ryzin [45] that if

$$\text{(V)} \qquad \int [D_N(x;S_N) - D_1(x,\theta)]^2 \, f_\xi(x)dx \to 0$$

in probability, where $f_\xi(x) = \xi_1 f_1(x) + \xi_2 f_2(x)$ , then the result III (in probability) holds. Van Ryzin [45] suggested a stochastic and recursive algorithm for estimating $D_1(x;\theta)$ for which (V) holds under fairly general conditions. His algorithm essentially involves window-kernels for estimating the density functions. Van Ryzin's work was inspired by the work of Aizerman et al [1] who proved (V) with recursive estimates using essentially method (i) along with an additional assumption that $D_1(x;\theta)$ is a finite linear combination of some known orthonormal functions in $L_2$ . For a generalization of the above work, see Gyorfi [17]. Wolverton and Wagner [51] proved that

(VI) $$\int [D_N(x;S_N) - D(x,\theta)]^2 \, dx \to 0$$

in prob./a.s. implies the result III in prob./a.s., when $f_i$'s are uniformly continuous (on $R^m$ ) . They suggested recursive estimates of $D(x,\theta)$ for which VI holds in probability when $f_i$'s are uniformly continuous and in a.s. when, in addition, $f_i$'s satisfy uniform Lipschitz condition. They also studied the rate of convergence when, specifically, $f_i$'s have bounded supports. Similar result on rate of convergence in probability was obtained by Rejtő and Revesz [33]. Watanabe [49,50] proved (VI) in prob. and a.s. along with their rates using recursive estimates for $D(x,\theta)$ following the method (v). Similar problem was studied by Tanaka [41] when the training sample constitutes dependent observations. Pelto [30] suggested to take the width of the window-kernel for estimating densities as the value which minimizes the deleted-counting estimate of the risk; however, his paper does not give any algorithm and his proofs are all based on heuristics.

It may be remarked that the rates of convergences derived in the papers cited above only reflect the performances of the suggested rules for future prediction, ignoring their performances in the past.

In passing, it may be noted that the estimate of density function by method (i) is not integrable; this estimate is further studied by Moore and Henrichson [27] and Wagner [47]. See also a recent paper by Wahaba [48].

3.3. Use of the two-sample test procedures.

Let $(X_1,\ldots,X_{n_1})$ and $(Y_1,\ldots,Y_{n_2})$ be the observations on random samples from $\pi_1$ and $\pi_2$ respectively. Let $(Z_1,\ldots,Z_n)$ be the observations on the $n$ units to be classified. These units form a random sample either from $\pi_1$ or from $\pi_2$. Let $F_0$ be the common c.d.f. of $Z_i$'s. Then the problem is to test $F_0 = F_1$ vs. $F_0 = F_2$.

Two-sample test statistics are often used to devise rules for the above problem. The basic heuristic ideas can be described as follows.

(a) Use Z's and X's to test $F_0 = F_1$ vs. $F_0 = F_2$; let $T_1$ be a test statistic such that large values of $T_1$ lead to the rejection of $F_0 = F_1$. Similarly use Z's and Y's to test $F_0 = F_2$ vs. $F_0 = F_1$; let $T_2$ be a test statistic such that large values of $T_2$ lead to the rejection of $F_0 = F_2$. Then define a rule which accepts $F_0 = F_1$ if $T_1 < T_2$, and accepts $F_0 = F_2$ if $T_2 < T_1$. One may also compare the critical levels of $T_1$ and $T_2$ instead of comparing $T_1$ and $T_2$ directly.

(b) Assume $F_0 = F_1$ and treat Z's and X's as i.i.d. observations. Get an estimate of the divergence between $F_1$ and $F_2$ by using a test statistic for testing $F_1 = F_2$ vs. $F_1 \neq F_2$. Similarly assume Z's and

Y's as i.i.d. observations and determine the corresponding estimate of the divergence. A rule now can be devised by comparing these two estimates of divergence.

It is well-known that a distribution-free rule cannot be derived for the pattern recognition problem posed above. The performance of a rule can only be judged in specific situations except for some broad asymptotic results.

Das Gupta [ 9 ] used the idea (a) and suggested a rule based on Wilcoxon-statistic; he showed that such a rule is consistent (i.e., the error probabilities tend to 0 as $n,n_1,n_2 \to \infty$ ). Hudimoto [20] also used Wilcoxon-statistic when $F_1(x) \geq F_2(x)$ for all $x$ , and derived some bounds for the probabilities of error. Kinderman [23] used more or less the idea (b) in deriving rules based on rank-scores when $F_1(x) = F_2(x+\theta)$, $\theta > 0$ , and studied the asymptotic efficiencies of those rules in Pitman's sense. Chandra and Lee [ 7] suggested a rule which first uses Wilcoxon-test for $F_1(x) > F_2(x)$ vs. $F_1(x) < F_2(x)$ based on X's and Y's , and then tests $F_0 = F_1$ vs. $F_0 = F_2$ using another Wilcoxon-type statistic and the result of the first test. They studied the asyptotic properties of this rule as $n_1$ and $n_2$ tend to $\infty$ . It may be noted that this rule is asymptotically equivalent to Das Gupta's rule [ 9]. See Chatterjee (J. Multiv. Anal., 1973, 3, 26-56) for a related work.

A minimum distance rule can be defined following (a) above with $T_1$ and $T_2$ as distances between the respective empirical c.d.f.'s. Matusita [26] obtained bounds on the probabilities of error of such a rule for the discrete case with Matusita-distance. Das Gupta [9 ] proved the consistency (as $n,n_1,n_2 \to \infty$) of such rules and derived bounds for the probabilities of error using Kolmogorov-distance. No detailed studies on these rules are yet available.

For simplicity, consider the minimum distance rule with Kolmogorov-distance for $n = 1$. It can be shown that such a rule decides $F_o = F_1$ if $|r_1/n_1 - 1/2| < |r_2/n_2 - 1/2|$, and $F_o = F_2$ if $|r_1/n_1 - 1/2| > |r_2/n_2 - 1/2|$, where $r_1$ = # of $X_i$'s $< Z_1$ and $r_2$ = # of $Y_i$'s $< Z_1$. We have studied the probabilities of error of this rule, and the results will be reported elsewhere. In particular, suppose $F_1(x) = G(x-\theta_1)$, $F_2(x) = G(x-\theta_2)$, where $G$ is a continuous distribution, symmetric about 0. Then both the conditional probabilities of error tend to $G(-|\theta_1-\theta_2|/2)$ with probability 1 as $n_1, n_2 \to \infty$.

3.4 Use of Tolerance Regions.

The use of tolerance regions (and statistically equivalent blocks) in classification was suggested by Anderson [ 3 ]. The basic idea is quite related to the problem of estimating a density function, and in that form it appears in the work of Fix and Hodges [12]. Quesenberry and Gessaman [31] suggested a method for contructing a rule based on tolerance regions which is asymptotically optimal; however, their idea is not very useful since the construction of such a rule depends on some inherent known structures of the distributions. Later, Anderson and Benning [ 2 ] and Beakley and Tuteur [ 4 ] suggested some other heuristic models. So far no theoretical results are available on these rules. This is due to the fact that very little is known on the performance of a tolerance region under a different distribution. Gessaman and Gessaman [14] studied some of these rules by Monte Carlo method.

3.5 Empirical Best-of-Class Rules.

The basic idea can be described as follows: Consider a class $\Delta$ of

rules, and let $C(\delta)$ be the proportions of correct identifications when $\delta \epsilon \Delta$ is applied to identify the observations in the training sample. Let $\delta_N \epsilon \Delta$ be a rule which maximizes $C(\delta)$ in $\Delta$. Then $\delta_N$ is called an empirical best $\Delta$-class rule. Let $\delta^* \epsilon \Delta$ be a rule which maximizes the probability of correct classification in the class $\Delta$.

Stoller [40] proved the convergence (in prob.) of the conditional risk of $\delta_N$ to the risk of $\delta^*$ in the univariate case when $\Delta$ is the class of rules determined by single cutoff points. Glick studied the convergence (a.s.) of $C(\delta_N)$ to the probability of correct classification of $\delta^*$ with special emphasis on the class of linear rules. The other papers in this area can be found in Duda and Hart (Pattern Classification and Scene Analysis, 1973, Wiley), although these are not of much statistical interest.

General results regarding the existence of $\delta_N$ and algorithm to determine $\delta_N$ are not yet available (except in the case considered by Stoller). General asymptotic results easily follow from the known asymptotic properties of empirical c.d.f.'s.

We suggest the following rule in the multivariate case, which is easy to apply. First treat the problem separately for each variate. However this will lead to some inconsistent decisions or indecision zones. Each of these zones can then be treated successively and separately by each variate to successively reduce the number of such indecision zones. It is believed that this rule will be asymptotically optimal for a large number of classes.

3.6 Rank-Distance Rules.

The basic idea is to find distances of the observations in the training sample from the observation $X$ to be classified and construct rules based on

the ranks of these distances and the corresponding pattern-class numbers.

Fix and Hodges [12] suggested the following rule $\delta_N$, termed as 1-NN rule: Classify $X$ to the pattern-class of the nearest neighbor (NN) of $X$. It can be shown (possibly given in [12] ) that under mild conditions, the probability of misclassification

$$\alpha_1(\delta_N;f) \to \int \frac{p_2 f_2(x)\, f_1(x)\, dx}{p_1 f_1(x) + p_2 f_1(x)},$$

where $p_i = \lim n_i/(n_1 + n_2)$, as $n_1, n_2$ (the sample sizes from $\pi_1$ and $\pi_2$) tend to $\infty$. For the mixed population case, Cover and Hart [ 8 ] have shown that

$$R(\delta_N;\theta) \to \int_{R^m} \frac{2\xi_1\xi_2 f_i(x)\, f_2(x)\, dx}{\xi_1 f_1(x) + \xi_2 f_2(x)} \equiv R_o$$

as $N \to \infty$, when the sample space of each $X_i$ is $R^m$ (or slightly more general). Wagner [46] has shown that under very mild conditions the conditional risk of $\delta_N$ tends to $R_o$ in probability (in a.s. under additional restrictions).

Fix and Hodges [12] also suggested the $K_N$-NN rule which can be described as follows. Let $M_{Ni}$ be the number of observations in the training sample with the pattern class $i$ that belong to the $K_N$ nearest neighbors of $X$. Then the $K_N$-NN rule decides the pattern-class of $X$ as 1, if $M_{N1}/n_1 > M_{N2}/n_2$, where $n_i$ is the number of observations in the training sample with the pattern-class $i$. One may also consider a rule by comparing $M_{N1}$ and $M_{N2}$. Johns [21] stated that under mild conditions the risk of the later rule tends to the risk of $\delta^*$ (Bayes), when $K_N \to \infty$,

$K_N/N \to 0$. The convergence of the conditional risk of this rule to the risk of $\delta^*$ (in different modes) can be obtained from the results in 3.1. Some other theoretical results on $K_N$-NN rule are claimed in the literature, although they were not proved with rigor.

It may be noted that NN rules are also related to the rules based on estimates of density functions. All the papers in this area deal only with the problem of classifying one unit.

3.7 Conclusion.

Many nonparametric rules are suggested in the literature from heuristic viewpoints. Asymptotic properties of most of these rules are not difficult to obtain, although good studies on the rates of convergences and asymptotic expansions of risks would be more useful. The usefulness of a rule is determined by its simplicity, as well as, by its robustness. Studies on robustness and small-sample behavior of these rules are quite limited. The relative comparisons (finite-sample or asymptotic) of some of the popular rules in specific situations are also called for.

## 4. Sequential Rules.

The sequential pattern recognition may involve sequential experimentation and sequential decision rules. A sequential experiment may arise in any combination of the following three situations. (a) Selection of components of the measurement vector on each unit in the training sample, as well as, in the sample of units to be classified. (b) Selection of the sample size of the training sample. (c) Selection of the sample size of the units to be classified when all the units are known to belong to one population. The basic object for using a sequential rule is to attain prescribed probabilities of errors and to reduce the average sample size; in Bayes' formulation

(involving probabilities of errors and cost of observations) the object is to reduce Bayes' risk. When the pattern-class densities are known, a Bayes rule can easily be derived following Wald's formulation; however, when the densities are unknown the problem is too involved and no satisfactory results are yet available.

Following the ideas of Hoeffding and Wolfowitz [19] , Das Gupta and Kinderman [11] introduced an important notion termed as "classifiability" for the situations (b) and (c) above. Consider three independent random vectors (of the same dimension) $X_0$, $X_1$ and $X_2$, where the c.d.f. of $X_i$ is $F_i$. It is known that $F$ is either equal to $F_1$ or $F_2$, and $F_1 \times F_2$ lies in a given set $\Omega$. The problem is to decide $F_0 = F_1$ or $F_0 = F_2$ based on a sequence of observations on $(X_0, X_1, X_2)$. This problem is said to be sequentially (finitely) classifiable, if for every $\alpha (0<\alpha<1)$ there exists a sequential rule (finite sample size rule) which terminates with probability 1 such that the probabilities of that rule are uniformly (in $\Omega$) less than $\alpha$. Necessary and sufficient conditions for sequential and finite classifiability are given by Das Gupta and Kinderman [11]. With the object of controlling the error probabilities uniformly and arbitrarily it is also important to find out whether it is necessary and sufficient to get observations only on $X_1$ (or on $X_2$) or on both $X_1$ and $X_2$. This problem is analysed also in Das Gupta and Kinderman [11]. In particular, suppose $F_1 = N_p(\mu_1, \Sigma)$, $F_2 = N_p(\mu_2, \Sigma)$. Then the problem is sequentially classifiable based on observations on $(X_0, X_1, X_2)$ if, and only if, $\mu_1 \neq \mu_2$. The problem is sequentially classifiable or finitely classifiable based on observations on $(X_0, X_1)$ if $\inf_{\Omega} \|\mu_1 - \mu_2\| > 0$, or $\inf_{\Omega} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) > 0$, respectively.

Following Hoeffding and Wolfowitz [19] we present a "minimum distance" sequential rule. Consider $n$ observations on $(X_0,X_1,X_2)$, and let $F_i^{(n)}$ be the empirical c.d.f. based on $n$ observations on $X_i$ $(i=0,1,2)$. Let $\Omega \subset \mathfrak{F} \times \mathfrak{F}$, where $\mathfrak{F}$ is a class of distributions on the space of $X_i$, and let $d$ be a uniform consistent distance function defined on $\mathfrak{F} \times \mathfrak{F}$. (We also assume that $d$ is defined on empirical c.d.f.'s). Assume that

$$\Omega_d = \{F_1 \times F_2 \in \Omega: \ d(F_1, F_2) = 0\}$$

is null. Let $\{C_i\}$ be a sequence of positive constants decreasing to 0, and $\{N_i\}$ be a sequence of strictly increasing positive integers. Now we define the rule as follows. (See [23]).

Take samples of sizes $n_1$, $n_2-n_1,\ldots,$ until $\Delta_i \geq C_i$, where $\Delta_i = \max\{d(F_0^{(ni)},F_1^{(ni)}), d(F_0^{(ni)},F_2^{(ni)})\}$. Setting $N = n_i$ make the terminal decision as follows. Decide $F_0 = F_1$ iff $d(F_0^{(N)},F_1^{(N)}) < d(F_0^{(N)},F_2^{(N)})$.

Given $\alpha$, the sequences $\{n_i\}$ and $\{C_i\}$ can be chosen such that $P(N < \infty) = 1$ and the probabilities of errors are less than $\alpha$. If $d$ is Kolmogorov-distance then $EN < \infty$ besides the above. However, the distribution of $N$ needs further study. For the two-population problem, Kurz and Woinsky [24] suggested a nonparametric sequential rule based on Wilcoxon statistic considering the situations (b) and (c), when $\int F_2(x)dF_1(x) < 1/2$ or $F_2(x) = F_1(x-\theta)$ with known $\theta$. They considered asymptotic properties of their rules as the maximum of the error probabilities (denoted by $\alpha$) tend to 0. Following the technique of Chow-Robbins they proved the asymptotic efficiency of the sample size, although the result is not very meaningful.

Moreover, they proved that the difference between the maximum of the error probabilities of their rule and $\alpha$ tends to 0 as $\alpha$ tends 0; such a result is almost trivial and throws no light at all on the performance of their rule. One should consider the limit of the ratio of the two above instead of their difference.

Srivastava [39] proposed sequential rules when $F_1 = N_p(\mu_1,\Sigma)$ and $F_2 = N_p(\mu_2,\Sigma)$ in the following two cases: (i) $\mu_1-\mu_2$ known but $\Sigma$ unknown; (ii) Both $\mu_1-\mu_2$ and $\Sigma$ are unknown. Given $\alpha$, he constructed a sequential rule for the case (i) such that the error probabilities of the rule are less than $\alpha$ and its sample size is asymptotically efficient (in comparison to the sample size when the parameters are known and as $(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1-\mu_2)\to 0$). For the case (ii), he showed that the limits of the error probabilities of his rule are $\leq \alpha$. Srivastava followed the ideas of Chow-Robbins and Simons [37]; however, his proof for the case (i) is incomplete and it is wrong for the case (ii). The main error lies in the fact that the notion of a.s. convergence as $(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1-\mu_2)\to 0$ is not well defined.

For the situation (a) described above, there are many apparently good results available in the literature, especially in Fu's book [13]. Unfortunately, most of the results are blind copies of the two-sample sequential rules and they lack sufficient rigor, as well as, meaningful formulation. Some heuristic rules, as in Smith and Yau [38], may be studied further with proper rigor.

Practically very few interesting results are available in the study of sequential rules. The problem requires first a meaningful formulation and a useful definition for asymptotic efficiency. For example in Srivastava's case (ii) no asymptotically efficient sequential rule would exist and one may

then focus on the loss of efficiency. It seems that the problem may well be studied from Chernoff's viewpoint after introducing sampling cost.

References.


[ 1]. Aizerman, M.A., Braverman, E.M. and Rozonoer, L.T. (1964). The probability problem of pattern recognition learning and the method of potential function. Automat. and Remote Control, 26, 1175-1190.

[ 2]. Anderson, M.W. and Benning, R.D. (1970). A distribution-free discrimination procedure based on clustering. I.E.E.E Trans. Inform. Theory, IT-16, 541-548.

[ 3]. Anderson, T.W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. Proc. 1st Internat. Symp. Multiv. Anal., Ed. P. R. Krishnaiah. Academic Press, New York. pp. 5-27.

[ 4]. Beakley, G. W. and Tuteur, F. B. (1972). Distribution-free pattern verification using statistically equivalent blocks. I.E.E.E Trans. Computers, C-21, 1337-1347.

[ 5]. Cacoullos, T. (1966). Estimation of a multivariate density. Ann. Inst. Statist. Math.

[ 6]. Čencov, N.N. (1962). Evaluation of an unknown distribution density from observations. Soviet Mathematics, 3, 1559-1562.

[ 7]. Chanda, K.C. and Lee, J.C. (1975). A class of nonparametric classification rules. To appear in Some Statistical Methods Useful in Oil Exploration. Ed. D. B. Owen.

[ 8]. Cover, T.M. and Hart, P.E. (1967). Nearest neighbor pattern classification. I.E.E.E. Trans. Inform. Theory, IT-16, 26-31.

[ 9]. Das Gupta, S. (1964). Nonparametric classification rules. Sankhya, A, 26, 25-30.

[10]. Das Gupta, S. (1973). Theories and methods in classification: A review. Discriminant Analysis and Prediction. Ed. T. Cacoullos. Academic Press, New York. pp. 77-137.

[11]. Das Gupta, S. and Kinderman, A. (1974). Classifiability and designs for sampling. Sankhya, A, 36, 237-250.

[12]. Fix, E. and Hodges, J.L. (1951). Nonparametric discrimination: Consistency properties. U. S. Air Force, School of Aviation Medicine, Report No. 4. Randolph Field, Texas.

[13]. Fu, K.S. (1968). Sequential methods in pattern recognition and machine learning. Academic Press, New York.

[14]. Gessaman, M.P. and Gessaman, P.H. (1972). A comparison of some multivariate discrimination procedures. Jour. Amer. Statist. Assoc., 67, 468, 472.

[15]. Glick, N. (1969). Estimating unconditional probabilities of correct classification. Tech. Report No. 3. Stanford Univ., Department of Statistics, Stanford.

[16]. Glick, N. (1972). Sample-based classification procedured derived from density estimators. Jour. Amer. Statist. Assoc., 67, 116-122.

[17]. Győrfi, L. (1972). Estimation of a probability density and optimal decision functions in reproducing kernel Hilbert space. Progress in Statistics, Vol. I. Ed. J. Gani et.al. North-Holland Publishing Co., London.

[18]. Hannan, J.F. and Robbins, H. (1955). Asymptotic solution of the compound decision problem for two completely specified distributions. Ann. Math. Statist., 26, 37-51.

[19]. Hoeffding, W. and Wolfowitz, J. (1958). Distinguishability of probability measures. Ann. Math. Statist., 29, 700-718.

[20]. Hudimoto, H. (1964). On a distribution-free two-way classification. Ann. Inst. Statist. Math., 16, 247-253.

[21]. Johns, M.V. (1967). An empirical Bayes approach to nonparametric two-way classification. Studies in Item Analysis and Prediction. Ed. H. Solomon. Stanford University Press, Stanford.

[22]. Kanal, L. (1974). Patterns in Pattern Recognition: 1968-1974. I.E.E.E. Trans. Inform. Theory, IT-20, No. 6, 697-722.

[23]. Kinderman, A. (1972). On some properties of classification: classifiability, asymptotic relative efficiency and a complete class theorem. Tech. Report No. 178, Dept. of Statistics, University of Minnesota, Minneapolis.

[24]. Kurz, L. and Woinsky, M.M. (1969). Sequential nonparametric two-way classification with prescribed maximum asymptotic error. Ann. Math. Statist., 40, 445-455.

[25]. Loftsgaarden, D.O. and Quesenberry, C.P. (1965). A nonparametric estimate of a multivariate density estimation. Ann. Math. Statist., 36, 1049-1051.

[26]. Matusita, K. (1956). Decision rule, based on the distance, for the classification problem. Ann. Inst. Statist. Math., 8, 67-77.

[27]. Moore, D.S. and Henrichson, E.G. (1969). Uniform consistency of some estimates of a density function. Ann. Math. Statist., 40, 1499-1502.

[28]. Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist., 33, 1065-1076.

[29]. Patrick, E.A. (1972). Fundamentals of Pattern Recognition. Prentice-Hall, Englewood Cliffs, New Jersey.

[30]. Pelto, C.R. (1969). Adaptive nonparametric classification. Technometrics, 11, 775-792.

[31]. Quesenberry, C.P. and Gessaman, M.P. (1968). Nonparametric discrimination using tolerance regions. Ann. Math. Statist., 39, 664-673.

[32]. Rao, C.R. (1952). Advanced Statistical Methods in Biometric Research. Wiley, New York.

[33]. Rejtő, L. and Revesz. (1973). Density estimation and pattern classification. Problems of Control and Inform. Theory, 2, 67-80.

[34]. Robbins, H. (1964). The empirical Bayes approach to statistical decision function. Ann. Math. Statist., 35, 1-20.

[35]. Samuel, E. (1963). Asymptotic solutions of the sequential compound decision problem. Ann. Math. Statist., 34, 1079-1094.

[36]. Samuel, E. (1963). Note on a sequential compound decision problem. Ann. Math. Statist., 34, 1095-1097.

[37]. Simons, G. (1968). On the cost of not knowing the variance when making a fixed-width confidence interval for the mean. Ann. Math. Statist., 39, 1946-1952.

[38]. Smith, S.E. and Yau, S.S. (1972). Linear sequential pattern classification. I.E.E.E. Trans. Inf. Theory, 673-678.

[39]. Srivastava, M.S. (1973). A sequential approach to classification: Cost of not knowning the covariance matrix. Jour. Multiv. Anal., 3, 173-183.

[40]. Stoller, D.C. (1954). Univariate two-population distribution-free discrimination. Jour. Amer. Statist. Assoc., 49, 770-777.

[41]. Tanaka, K. (1970). On the pattern classification problems by learning, (I and II). Bull. Math. Statist., 14, 31-49 and 61-73.

[42]. Van Ryzin, J. (1965). Nonparametric Bayesian decision procedure for (pattern) classification with stochastic learning. Proc. IV Prague Conf. on Inf. Theory, Statist. Dec. Functions, and Random Processes. 479-494.

[43]. Van Ryzin, J. (1966). Repetitive play on finite statistical games with unknown distributions. Ann. Math. Statist., 37, 976-994.

[44]. Van Ryzin, J. (1966). Bayes risk consistency of classification procedures using density estimation. Sankhya, A, 28, 201-270.

[45]. Van Ryzin, J. (1967). A stochastic a posteriori updating algorithm for pattern recognition. Math. Anal. Appl., 20, 359-379.

[46]. Wagner, T. J. (1971). Convergence of the nearest neighbor rule. I.E.E.E. Trans. Inf. Theory, IT-17, No. 5, 566-571.

[47]. Wagner, T.J. (1973). Strong consistency of a nonparametric estimate of a density function. I.E.E.E. Trans. Systems Man Cybernetics, SMC-3.

[48]. Wahaba, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. Ann. Statist., 3, 15-29.

[49]. Watanabe, M. (1972). On asymptotically optimal algorithms for pattern classification problems. Bull. Math. Statist., 15, 31-48.

[50]. Watanabe, M. (1974). On convergences of asymptotically optimal discriminant function for pattern classification problems. Bull. Math. Statist., 16, 23-34.

[51]. Wolverton, C.T. and Wagner, T.J. (1969). Asymptotically optimal discriminant functions for pattern classification. I.E.E.E. Trans. Inf. Theory, IT-16, No. 2, 258-265.

[52]. Yamato, H. (1971). Sequential estimation of a continuous probability density function. Bull. Math. Statist., 14, 1-12.

[53]. Yao, J. (1972). A sequential classification into one of several populations. Bull. Math. Statist., 15, 19-28.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

# REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 258
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): Some Problems in Statistical Pattern Recognition
5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Somesh Das Gupta
8. CONTRACT OR GRANT NUMBER(s): DAAG-29-76-G-0038

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Department of Theoretical Statistics
University of Minnesota
Minneapolis, Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS:
US Army Research Office
Mathematics Division
P.O. Box 12211, Research Triangle Park, NC 27709

12. REPORT DATE: 20 January 1976
13. NUMBER OF PAGES: 23

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)
Reproduction in whole or in part is permitted for any purpose of the United States Government. Distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Pattern recognition; nonparametric rules; density estimates; tolerance regions rank-distance rules; sequential rules; classifiability.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
A review along with criticism of some recent work on nonparametric and sequential rules in statistical pattern recognition is given in this paper. Some new results and directions for future work are also discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)